LLM 모델 조사

현재 AI 시장에서 널리 사용되는 AI의 간략적인 요약은 다음과 같다

	LLaMA3.1	GPT-4	Claude	Gemini
학습 파라미터수	4050억개	수천억개	GPT 유사	유사
멀티모달	텍스트만	가능	텍스트 중심	가능
코딩능력	향상됨	높음	보통	보통
벤치마크	우수	우수	윤리적	효율적
목표	다목적	생성형AI	윤리적 인간친화적AI	기업친화적AI

위에서 언급하지 않은 LLM들과 요구사항에 맞는 모델들을 장점, 비용, 적합 이유를 기준으로 분석해보았다

오픈소스모델

<LLaMA 2 13B>

- 고품질 데이터로 파인튜닝 가능하며, 다양한 도메인에 적합. 컴퓨팅 자원은 GPU 클러스터 필요,

장점

- LLaMA 2 13B는 instruction fine-tuning을 통해 특정 도메인에 맞게 조정할 수 있으며, 고품질 데이터셋을 활용하여 네트워크 분석, 데이터 분석 등 다양한 작업에 적합하다.
- 상대적으로 낮은 비용으로 파인튜닝이 가능하다.

컴퓨팅 자원

- GPU 클러스터가 필요하며, 이는 상당한 수준의 컴퓨팅 자원을 요구한다.

적합 이유

- 다양한 기능과 높은 성능을 제공하며, 특히 중소기업의 PR 기능 향상에 필요한 텍스트 생성 및 분석 능력을 갖추고 있다.

<DistilBERT>
- 경량화된 모델로 적은 자원으로도 파인튜닝 가능. 비용 효율적

장점
- 경량화된 모델로, 작은 규모의 데이터셋에서도 효율적인 학습이 가능하다.
- DistilBERT는 비용 효율적인 모델로, 파인튜닝 비용이 상대적으로 낮다.

컴퓨팅 자원
- 적은 자원으로도 파인튜닝이 가능하여, 컴퓨팅 자원 제약이 있는 환경에 적합하다.

적합 이유
- PR 성공 및 실패 사례 구분, 텍스트 생성 등 기본적인 텍스트 분석 및 생성 작업에 적합하며, 특히 자원 제약이 있는 중소기업 환경에서 유용하다.

유료 모델

<GPT-4>
- 고성능 텍스트 생성 및 분석 가능. 높은 컴퓨팅 자원 필요, 비용은 상당히 높음.

장점
- PR 기능 향상을 위한 다양한 텍스트 생성 및 분석 작업에 적합하며, 특히 고품질 결과물을 요구할 때 유용하다.

컴퓨팅 자원
- 높은 GPU 클러스터 수준이 필요하며 일반적으로 1,000 토큰당 약 0.03~0.12달러의 비용이 발생합니다. 한 달에 1,000,000 토큰을 사용한다고 가정하면, 대략 40,000~160,000원 정도의 비용이 발생할 수 있다.

적합 이유
- 기본적인 텍스트 분석 및 생성 작업에 적합하며, 고품질의 창의적인 PR 컨텐츠 생성이 가능하다.

<Claude>

- 인간 친화적 모델로 고성능 텍스트 분석 및 생성 가능, 높은 컴퓨터 자원 필요

장점
- 사용자 요구에 맞춘 텍스트 생성에 강점
- 자연스러운 스타일, 창의적 문장 생성에 특화보도자료, 캐치프레이즈 등 PR 콘텐츠 작성에 적합
- 안전성, 해석 가능성, 인간 친화적 응답에 초점을 맞춘 모델로 긴 문서, 계약서 등도 한 번에 요약 가능

컴퓨팅 자원
- 중간 수준의 자원 필요, 비용은 GPT-4보다 낮음.

적합 이유
- 사용자 맞춤형 PR 콘텐츠 생성에 특화되어 있으며, 창의적인 캐치프레이즈 및 보도 기사 작성에 적합하다.

분석 결과

- 컴퓨터 자원과 비용측면으로 보았을 때, 오픈 소스 모델을 우선적으로 학습시킨다.

- 맥락 추론에 뛰어난 DistilBERT는 ㈜슈퍼히어로만의 분류기준을 통한 분류기능, 특정 도메인 학습의 강점을 지닌 LLaMA는 기업 데이터를 학습시켜 추천 기능을 우선으로 파인튜닝 성능을 측정하고 병합 모델의 경우도 고려해 볼 예정이다.

- 자원의 여유가 있다면 GPT-4 와 Claude를 사용해 PR에 적합하고 인간친화적인 모델 Claude의 성능을 확인하고 고품질 PR 컨텐츠 생성이 필요하면 GPT-4를 학습한 모델과 비교하며 연구를 할 계획이다.